Análisis comparativo de metodologías para la gestión de proyectos de minería de datos

Ing. Juan Miguel Moine[1], Dra. Silvia Gordillo[2], Dra. Ana Silvia Haedo[3]

1. Grupo de Investigación en Minería de Datos, UTN Rosario
2. Facultad de Informática, Universidad Nacional de La Plata
3. Facultad de Ciencias Exactas, Universidad Nacional de Buenos Aires

[1]juanmiguelmoine@gmail.com, [2]gordillo@lifia.info.unlp.edu.ar, [3]ahaedo@dc.uba.ar

Resumen. La sistematización del proceso de minería de datos es un punto importante para la planificación y ejecución de este tipo de proyecto. Algunas organizaciones implementan el modelo KDD, mientras que otras aplican un estándar más específico como CRISP-DM. Si la organización ha adquirido productos de la empresa SAS, tiene a su disposición una metodología especialmente desarrollada para los mismos, la metodología SEMMA. Por otro lado, la metodología Catalyst (conocida como P3TQ) está ganando cada vez mayor popularidad debido a su completitud y flexibilidad para adaptarse en distintos escenarios. En el presente trabajo se realiza un análisis comparativo de las diferentes metodologías vigentes para minería de datos, evaluando no sólo la estructura del proceso, sino también aspectos importantes para la gestión del proyecto.

Palabras clave: Minería de Datos, Gestión de Proyectos, Knowledge Discovery in Databases, Explotación de Información, CRISP-DM, SEMMA, Catalyst, P3TQ, Metodologías en Minería de Datos.

1 Introducción

La minería de datos es una disciplina que ha crecido enormemente en los últimos años. Las organizaciones han comprendido que los grandes volúmenes de datos que residen en sus sistemas pueden ser analizados y explotados para obtener nuevo conocimiento a partir de los mismos.

Minería de Datos o Explotación de Información, es el proceso de extraer conocimiento útil, comprensible y novedoso de grandes volúmenes de datos, siendo su principal objetivo encontrar información oculta o implícita, que no es posible obtener mediante métodos estadísticos convencionales. El proceso de minería se basa en el análisis de registros provenientes de bases de datos operacionales o bien bodegas de datos (Datawarehouse).

Los esfuerzos en el área de la minería de datos se han centrado en su gran mayoría en la investigación de técnicas para la explotación de información y extracción de patrones (tales como árboles de decisión, análisis de conglomerados y reglas de asociación). Sin embargo, se ha profundizado en menor medida el hecho de cómo ejecutar este proceso hasta obtener el "nuevo conocimiento", es decir, en las

metodologías. Las metodologías permiten llevar a cabo el proceso de minería de datos en forma sistemática y no trivial. Ayudan a las organizaciones a entender el proceso de descubrimiento de conocimiento y proveen una guía para la planificación y ejecución de los proyectos.

Algunos modelos conocidos como metodologías son en realidad un modelo de proceso: un conjunto de actividades y tareas organizadas para llevar a cabo un trabajo. La diferencia fundamental entre metodología y modelo de proceso radica en que el modelo de proceso establece qué hacer, y la metodología especifica cómo hacerlo. Una metodología no solo define las fases de un proceso sino también las tareas que deberían realizarse y cómo llevar a cabo las mismas.

En los inicios del año 1996, el modelo **KDD** [1] (Knowledge Discovery in Databases) constituyó el primer modelo aceptado en la comunidad científica que estableció las etapas principales de un proyecto de explotación de información. En su versión completa, KDD está formado por nueve etapas, donde la primera es el entendimiento del negocio. Formalmente el modelo establece que la minería de datos es la etapa dentro del proceso en la cual se realiza la extracción de patrones a partir de los datos. Sin embargo actualmente, en la comunidad científica y en la literatura, el término KDD y minería de datos se utilizan indistintamente para hacer referencia al proceso completo de descubrimiento de conocimiento.

A partir del año 2000, con el gran crecimiento en el área de la minería de datos, surgen tres nuevos modelos que plantean un enfoque sistemático para llevar a cabo el proceso: SEMMA [2], CRISP-DM [3] y Catalyst [4] (conocida como P3TQ). CRISP-DM se ha convertido en la metodología más utilizada, según un estudio publicado en el año 2007 por la comunidad KDnuggets (Data Mining Community's Top Resource).

SEMMA, desarrollada por el SAS Institute, se define como "el proceso de selección, exploración y modelado de grandes volúmenes de datos para descubrir patrones de negocio desconocidos". El nombre de esta terminología es el acrónimo correspondiente a las cinco fases básicas del proceso: Sample (Muestreo), Explore (Exploración), Modify (Modificación), Model (Modelado), Assess (Valoración).

La metodología SEMMA se encuentra enfocada especialmente en aspectos técnicos, excluyendo actividades de análisis y comprensión del problema que se está abordando. Fue propuesta especialmente para trabajar con el software de la compañía SAS. Este producto organiza sus herramientas (llamadas "nodos") en base a las distintas fases que componen la metodología. Es decir, el software proporciona un conjunto de herramientas especiales para la etapa de muestreo, otras para la etapa de exploración, y así sucesivamente. Sin embargo, el usuario podría hacer uso del mismo siguiendo cualquier otra metodología (CRISP-DM por ejemplo).

La metodología **Catalyst**, conocida como P3TQ (Product, Place, Price, Time, Quantity), fue propuesta por Dorian Pyle en el año 2003. Esta metodología plantea la formulación de dos modelos: el Modelo de Negocio y el Modelo de Explotación de Información.

El Modelo de Negocio (MII), proporciona una guía de pasos para identificar un problema (o la oportunidad del mismo) y los requerimientos reales de la organización. Contempla diferentes ámbitos para el proyecto de minería de datos, explicitando acciones específicas según el escenario desde el cual se parte. Para proyectos donde el problema u oportunidad de negocio no está definido, se recomienda comenzar analizando las relaciones P3TQ que existen en la cadena de

valor organizacional, es decir, aquellas relaciones precio/lugar/producto /tiempo/cantidad que son importantes para la empresa.

El Modelo de Explotación de Información (MIII), proporciona una guía de pasos para la construcción y ejecución de modelos de minería de datos a partir del Modelo de Negocio (MII).

El foco que propone la metodología Catalyst en su Modelo de Negocio sobre la cadena de valor organizacional, hizo que sea difundida en la comunidad científica como metodología "P3TQ", aunque ésta no sea su denominación original.

La metodología Catalyst, en sus dos modelos, está compuesta por una serie de pasos llamados "boxes". El concepto es que luego de llevar a cabo una acción, se deben evaluar los resultados y determinar cuál es el próximo paso (box) a seguir. La secuencia y la interacción entre los distintos pasos permiten una flexibilidad muy grande, y una amplia variedad de caminos posibles.

CRISP–DM, creada por el grupo de empresas SPSS, NCR y Daimer Chrysler en el año 2000, es actualmente la guía de referencia más utilizada en el desarrollo de proyectos de minería de datos. Estructura el proceso en seis fases: Comprensión del negocio, Comprensión de los datos, Preparación de los datos, Modelado, Evaluación e Implantación. La sucesión de fases, no es necesariamente rígida. Cada fase se descompone en varias tareas generales de segundo nivel. CRISP-DM establece un conjunto de tareas y actividades para cada fase del proyecto pero no especifica cómo llevarlas a cabo.

Objetivos. Analizar mediante un estudio descriptivo-comparativo las metodologías más difundidas en la actualidad para proyectos de minería de datos, evaluando los siguientes aspectos:

- El grado en que se incorporan actividades para la gestión del proyecto (como gestión del riesgo, de costos, de Recursos Humanos).
- El nivel de detalle de las tareas que componen cada fase, abriendo una discusión sobre qué modelos pueden ser realmente considerados una metodología.
- La viabilidad de cada modelo para la aplicación en diferentes escenarios (ya sea partiendo de un conjunto de datos o abordando una situación o problema organizacional).

2 Metodología

Para confrontar los modelos KDD, SEMMA, CRISP-DM y Catalyst, se realizará un análisis comparativo de los mismos considerando los siguientes aspectos:

a. Escenarios y puntos de partida considerados para el proyecto. Los proyectos de explotación de información pueden ser llevados a cabo en distintos escenarios. Según el punto de partida del proceso, es posible clasificarlos en:

- *Escenarios donde se aborda desde la minería de datos una situación organizacional* (un problema o una oportunidad), buscando patrones y relaciones que puedan colaborar con la misma.

- *Escenarios donde el proyecto comienza con un conjunto de datos* y el objetivo es explorarlos para encontrar relaciones interesantes que puedan ser útiles en el dominio de aplicación.

b. Estructura de las fases del proceso. Se analiza la estructura de cada modelo, en función de las siguientes fases generales comunes en los proyectos de minería de datos:

- Comprensión del negocio, evaluando el problema que se abordará y el contexto organizacional.
- Selección y preparación de los datos, limpieza y transformaciones necesarias para crear la vista minable.
- Aplicación de las técnicas de minería (análisis de regresión, árboles de decisión, redes neuronales, etc) y modelado de los nuevos patrones.
- Evaluación de los resultados obtenidos, analizando la posibilidad de implementarlos o bien de llevar a cabo nuevamente el proceso.
- Implementación y difusión del nuevo conocimiento dentro de la organización.

c. Nivel de detalle en las tareas de cada fase. Se evalúa el grado de profundidad con el que se describen las actividades y tareas en cada una de las fases del proceso. Algunos modelos describen sólo las fases generales, mientras que otros establecen las tareas específicas a llevar a cabo en cada una de ellas.

d. Actividades incorporadas para la gestión del proyecto. Los proyectos de minería de datos, al igual que en otras áreas como la Ingeniería del Software, requieren la ejecución de una serie de actividades que posibiliten el cumplimiento del objetivo del mismo. El PMBOK (Project Managment Body Of Knowdlege) es una colección de procesos y áreas de conocimiento generalmente aceptadas como las mejores prácticas dentro de la gestión de proyectos. El PMBOK es un estándar reconocido internacionalmente (IEEE Std 1490-2003) que provee los fundamentos de la gestión de proyectos que son aplicables a un amplio rango de proyectos, incluyendo construcción, software, ingeniería, etc. El PMBOK reconoce distintas áreas de conocimiento comunes a la mayoría de los proyectos. Entre ellas podemos destacar:

- *Gestión del Tiempo*: área de conocimiento que propone una serie de actividades cuyo objetivo es la conclusión en tiempo del proyecto. En éste área se incluye la estimación de la duración de las tareas y el desarrollo/control del cronograma del proyecto.
- *Gestión de Costos:* incluye los procesos involucrados en la planificación, estimación, preparación del presupuesto y control de costos, de forma que el proyecto se pueda completar dentro del presupuesto aprobado.
- *Gestión del Riesgo*: su objetivo es identificar, controlar y eliminar las fuentes de riesgo antes de que empiecen a afectar al cumplimiento de los objetivos del proyecto. Se busca disminuir la probabilidad y el impacto de los eventos adversos para el proyecto.
- *Gestión de Recursos Humanos*: se refiere a todos aquellos procesos que organizan y dirigen al equipo del proyecto. El equipo del proyecto está formado

por las personas a las que se le han asignado roles y responsabilidades para llevar adelante y concluir el proyecto.

- *Gestión del Alcance*: se refiere a la identificación de todas las tareas necesarias para completar el proyecto exitosamente. Cuando hablamos de alcance del proyecto, no nos referimos al "alcance del producto", sino al conjunto de tareas necesarias para entregar el producto. Una actividad frecuente en esta etapa es la creación de una WBS (Work Breakdown Structure) donde se detallan las tareas a menor nivel de detalle.

Las actividades que se llevan a cabo dentro de cada categoría pueden ser de *planificación* o bien de *control*. Las actividades de planificación incluyen la identificación de las tareas a realizar en el proyecto, estimación de la duración de las mismas, estimación de los recursos afectados y la definición del curso de acción. Las actividades de control tienen por objetivo el monitoreo del estado actual del proyecto para su comparación con lo planificado.

3 Resultados

a. Escenarios y puntos de partida considerados para el proyecto. Entre los cuatro modelos analizados, sólo SEMMA inicia el proyecto de minería a partir del conjunto de datos (la primera fase es el muestreo de los datos). CRISP-DM y KDD (en su versión completa de nueve pasos) comienzan con un análisis del negocio y del problema organizacional. Catalyst es la metodología más completa en este aspecto, ya que considera cinco escenarios posibles como punto de partida, entre los cuales se encuentra el inicio desde un problema u oportunidad de negocio.

b. Estructura de fases del proceso. KDD, CRISP-DM y Catalyst contemplan el análisis y comprensión del problema antes de comenzar el proceso de minería. SEMMA excluye esta actividad del modelo.
En todos los modelos se contempla la selección y preparación de los datos. Esta situación se repite para la fase de modelado, donde se aplican las técnicas de minería para obtener los nuevos patrones.
La fase de evaluación de los patrones obtenidos está presente también en todas las metodologías. En SEMMA, la evaluación e interpretación de estos patrones se realiza sobre el desempeño del modelo, mientras que en las otras metodologías la evaluación se realiza en función de la utilidad que se aporta al dominio de aplicación o problema organizacional.
La implementación de los resultados obtenidos es una fase que no está incluida en el modelo SEMMA. En CRISP-DM, se propone además una planificación para el control futuro y un análisis de cierre del proyecto (análisis postmortem). El análisis postmortem consiste en encontrar información objetiva acerca de la trayectoria de un proyecto, con la finalidad de poder hacer una evaluación abierta del equipo de trabajo, de las decisiones tomadas a lo largo del mismo, de las tecnologías empleadas y sus consecuencias, con el objetivo de incorporar lo aprendido en proyectos futuros.

Tabla 1. Fases del proceso de minería de datos en cada modelo

Fases	KDD	CRISP – DM	SEMMA	CATALYST
Análisis y comprensión del negocio	Comprensión del dominio de aplicación	Comprensión del negocio		Modelado del negocio
Selección y preparación de los datos	Crear el conjunto de datos Limpieza y pre-procesamiento de los datos Reducción y proyección de los datos	Entendimiento de los datos Preparación de los datos	Muestreo Comprensión Modificación	Preparación de los datos
Modelado	Determinar la tarea de minería Determinar el algoritmo de minería Minería de datos	Modelado	Modelado	Selección de herramientas y modelado inicial
Evaluación	Interpretación	Evaluación	Valoración	Refinamiento del modelo
Implementación	Utilización del nuevo conocimiento	Despliegue		Comunicación

c. Nivel de detalle en las tareas de cada fase

Los modelos KDD y SEMMA proponen sólo los pasos generales del proyecto de minería de datos, sin especificar puntualmente las tareas que deben llevarse a cabo en cada una de sus fases. En cambio, los modelos CRISP-DM y Catalyst, especifican con mayor detalle las actividades del proceso, aunque Catalyst señala además "cómo" realizarlas.

KDD y SEMMA se acercan más a un modelo de proceso que a una metodología, ya que sólo definen las fases generales. En proyectos donde se desee aplicar los mismos, cada organización deberá establecer las tareas y las actividades que implementará en cada etapa.

Si bien los modelos CRISP-DM y Catalyst no llegan a especificar con un alto nivel de detalle cómo realizar todas las tareas, podrían ser considerados una metodología ya que describen y puntualizan las actividades especificas a realizar en cada fase del proceso.

d. Actividades para la gestión del proyecto

En la tabla 2, podemos observar que tanto la metodología CRISP-DM como la metodología Catalyst proponen actividades de planificación para las distintas áreas de la gestión del proyecto, pero no explicitan tareas de control y monitoreo. KDD y SEMMA no incluyen actividades de gestión del proyecto.

Tabla 2. Actividades para la gestión del proyecto en cada modelo

	KDD	CRISP – DM	SEMMA	CATALYST
Gestión del alcance		Planificación del alcance en la tarea 1.4.		Planificación del alcance en la tarea 6 del Modelado del Negocio.
Gestión del tiempo		Planificación del tiempo en la tarea 1.4.		Planificación del tiempo en la tarea 6 del Modelado del Negocio.
Gestión del costo		Planificación del costo en la tarea 1.4.		Planificación del costo en la tarea 6 del Modelado del Negocio.
Gestión del riesgo		Gestión del riesgo en la tarea 1.2.		Planificación del riesgo en la tarea 8 del Modelado del Negocio.
Gestión de los recursos humanos		Planificación de recursos humanos en la tarea 1.4.		Planificación de recursos humanos en la tarea 6 del Modelado del Negocio.

En CRISP-DM las actividades de planificación se ven reflejadas en las tareas 1.2 (*Evaluación de la situación*) y 1.4 (*Crear un plan para el proyecto de minería de datos*). Si bien no se explicitan tareas de seguimiento y control, en el modelo se aclara que el plan del proyecto debe ser revisado (y de ser necesario modificado), antes de comenzar con cada fase del proceso.

En Catalyst, las actividades de planificación se llevan a cabo en la Metodología para el Modelado del Negocio (MII). Tomando como escenario de partida un problema u oportunidad organizacional, la planificación del alcance, tiempo, costo y recursos humanos se proponen en la tarea 6, *"Armar el caso de negocio"*. La planificación del riesgo está presente en la actividad 8, *"Describir la situación del negocio para el proceso de minería"*. La metodología no explicita tareas de control y seguimiento del proyecto.

4 Conclusión

En este análisis se ha evidenciado la existencia de dos tipos de modelos para llevar a cabo el proceso de minería de datos.

Por un lado encontramos aquellos que están más cercanos a un modelo de proceso, ya que sólo proponen las fases generales para el proceso de minería de datos y no incorporan actividades para la gestión del proyecto. Estos modelos son KDD y SEMMA, los cuales no llegan a ser una metodología propiamente dicha y dejan a

criterio del equipo de trabajo la definición de las actividades a realizar en cada etapa del proyecto. Particularmente SEMMA excluye dos etapas importantes del proceso como son el análisis del negocio y la difusión del nuevo conocimiento, evidenciando que el modelo está orientado especialmente a aspectos técnicos.

Por otro lado, los modelos CRISP-DM y Catalyst podrían ser considerados una metodología, por el nivel de detalle con el que describen las tareas en cada fase del proceso, y porque incorporan actividades para la gestión del proyecto (como gestión del tiempo, costo, riesgo). En este aspecto, ninguno de los dos modelos incorpora actividades para el control y monitoreo del plan de trabajo.

La metodología Catalyst sobresale en su fase de Modelado del Negocio (MII), contemplando cinco puntos de partida para el proyecto, que finalmente conducirán a la definición de un conjunto de requerimientos y a una situación organizacional que deberá ser abordada desde la minería de datos.

Si hablamos entonces de metodologías para la gestión de un proyecto de minería de datos, los modelos a tener en cuenta deberían ser CRISP-DM y Catalyst, los cuales se encuentran en un nivel similar de completitud en la mayoría todos los aspectos evaluados.

Agradecimientos. Al programa de Becas Bicentenario de Investigación y Postgrado, Universidad Tecnológica Nacional.

Referencias

1. Fayyad, U.: Advances in Knowledge Discovery and Data Mining. *MIT Press* (1996).
2. SAS Institute: Data Mining and the Case for Sampling. *Data Mining Using SAS Enterprise Miner*. www.sasenterpriseminer.com/documents/SAS-SEMMA.pdf (1998) Acceso Julio 2010.
3. Chapman, P.; Clinton, J. y otros: CRISP-DM 1.0 Step by step guide. *SPSS*. www.crisp-dm.org/CRISPWP-0800.pdf (2000) Acceso Noviembre 2010.
4. Pyle, D.: Business Modeling and Data Mining. *Morgan Kaufmann Publishers* (2003).
5. Fayyad, U. y otros: The KDD process for extracting useful knowledge from volumes of data. *ACM vol. 39* (1996).
6. Azevedo, A.: KDD, SEMMA AND CRISP-DM a parallel overview. *AIDIS* (2008).
7. Britos, P.: Procesos de explotación de información basados en sistemas inteligentes. *Universidad Nacional de La Plata, Argentina* (2008).
8. Pollo-Cattaneo, F. y otros: Ingeniería de Proyectos de Explotación de Información. *WICC 2010. ISBN 978-950-34-0652-6* (2010).
9. Mariscal, G. y otros: A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review, Vol. 25:2, 137–166* (2010).